DOCUMENT RESUME

ED 186 446 TM 800 125

AUTHOR Myers, Charles T.
TITLE Test Length and Validity: An Application of Test Theory to a Finite World.
PUB DATE Mar 78
NOTE 14p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Item Analysis; Multiple Choice Tests; *Test Construction; Testing Problems; *Test Items; *Test Reliability; Test Theory; *Test Validity
IDENTIFIERS *Test Length

ABSTRACT

The viewpoint is expressed that adding to test reliability by either selecting a more homogeneous set of items, restricting the range of item difficulty as closely as possible to the most efficient level, or increasing the number of items will not add to test validity and that there is considerable danger that efforts to increase reliability may reduce validity. It is also stated that short tests may have useful validity if the items are well made and carefully selected. More stress should be placed on item quality and less on quantity. It is recommended that occasional well-funded item validity studies be initiated. (BH)

****
* Reproductions supplied by EDPS are the best that can be made *
* from the original document. *
****

ED186446

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TEST LENGTH AND VALIDITY

AN APPLICATION OF TEST THEORY TO A FINITE WORLD

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C. Myers

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Charles T. Myers

Educational Testing Service

Princeton, New Jersey

March 1978

TM 800125

TEST LENGTH AND VALIDITY

AN APPLICATION OF TEST THEORY TO A FINITE WORLD

A widely accepted and commonly practiced axiom of the academic testing business states that, if a test is to be used in making important decisions the test must be reliable. Let us consider a situation that demands an important decision. Suppose you are a student in a school for parachute jumping. Part of the curriculum is learning to fold your own parachute; and, before you progress to actually jumping from a plane, you must satisfactorily complete the parachute folding part of the course. It is unlikely that formal written tests are used for this purpose, but it is obvious that some sort of assessment must be made. Now if I were a student in such a school, the reliability of this assessment would not be my primary concern. Instead I would be concerned to know whether that assessment was valid! If there is a textbook on parachute folding, I think I could write an acceptably reliable multiple-choice test on parachute folding; but would you accept a high score on that test as evidence that you were ready to go up in a plane and start jumping?

There is not much point in my saying what I have to say today unless it can be agreed that it is quite possible for the validity of an achievement test to be different from its reliability. This is the central issue of what I have to say and it is, at least in part, a matter of definition.

For this purpose the word reliability will mean what that word means in Kuder and Richardson's well known formula number twenty. The term validity will mean the zero-order product-moment correlation between an achievement test and the best available approximation of its ideal criterion for the same sample of persons on whom the test's reliability was computed, with the further specification that the criterion may not be another multiple-choice test.

My thesis has three main points. The first is that an academic achievement test may be highly reliable and yet have an unsatisfactory level of validity; or that, conversely, a test may be more than usually valid and yet have only a modest level of reliability. Within the range of common experience in achievement testing, reliability and validity may be quite independent. The second point is that as far as the usefulness of a test is concerned, validity is all important and reliability is often irrelevant. The usefulness of a test must be judged by the proportion of decisions that are improved by the use of the test, or by the extent to which such decisions are improved. Whether or not a test has more reliability than is needed to achieve a given validity makes no difference. If you must choose between two tests that are equally valid, your decisions are not improved by choosing the more reliable test. The third and final point is that seeking reliability for its own sake may result in reducing the validity of a test. I suspect that, in fact, this happens frequently. I suspect that this is particularly likely to happen when either testing time or test development budget is limited. It seems particularly likely to happen when a limited budget results in item selection from a limited item pool.

Theoretical Analysis

Test-score theory provides test developers with three different ways to increase the reliability of a test. One way is to select a more homogeneous set of items. A second way is to restrict the range of item difficulty as closely as possible to the most efficient level. (Perhaps the best way to do this is to use a system of "tailored testing" in which the choice of items used in each test depends upon the ability level of each candidate; but for now I will restrict my comments to such group concepts as reliability, validity, and the traditional variety of standardized test.) The third way to increase test reliability is to increase the number of items.

Test score theory may suggest other ways of increasing reliability, but I am sure these three are the most familiar and probably the most effective. The option of giving the test to a more heterogeneous group is not in the same class of strategies and is not relevant to this discussion.

Let us now consider what these three strategies for increasing test reliability are likely to do to test validity. It is well known, but frequently disregarded, that the best way to maximize the validity of a test battery is to select tests that are individually valid but which have low intercorrelations among themselves. This is just as true of the selection of test items as it is of the selection of tests. Hence maximizing the homogeneity of a selected set of items tends to reduce the validity of the set. This point has been made repeatedly for many years, but apparently it has had only a modest effect on test

development policies. The second way to increase test reliability is to restrict the range of item difficulties. I cannot tell you anything about the theoretical effect this should have on test validity and I know of only one experimental study of this question. The study used an ex post facto design, and the results with respect to validity did not quite reach the five percent level of statistical significance. However, the average validity coefficient for the items with a wide spread of difficulties was slightly higher than was the validity of the more reliable set of items with a narrow spread of item difficulties. I am fairly confident that, in this situation, restricting the range of item difficulties did not improve the validities.

Classical test-score theory provides a clear and simple statement about the relationship of test length to both test reliability and test validity. In simplest terms, other things being equal, increasing the length of a test by adding to the number of items increases both reliability and validity. But the effect of increasing the number of items tends to be much greater in increasing reliability than it is in increasing validity.

Gulliksen (1951) provides a formula for test reliability (r_{tt}) as a function of test length (N) and of average item reliability ($\bar{r}_{ii}$) which may be written

$$r_{tt} = \bar{r}_{ii}(N/(1 + (N - 1)\bar{r}_{ii})).$$

The formula for test validity (r_{tc}) is very similar in appearance, but it includes one new term, average item validity ($\bar{r}_{ic}$), and may be written

$$r_{tc} = \bar{r}_{ic}\sqrt{N/(1 + (N-1)\bar{r}_{ii})} \quad .$$

The term under the square root sign in the second formula is identical with the term in parentheses in the first formula.

In order to appreciate the relationship between test length and reliability and validity, it may be useful to consider some hypothetical numerical examples. Since the reliability or validity of a single item is not often reported, it may be instructive to estimate what values may be typical. By use of the Spearman-Brown formula it can be computed that if a 50-item test has a reliability of .90, then the average item in that test will have a reliability of .15. (For many test programs these values would be well above the average, but they are within the range of practical possibility.) Test validities range from zero or negative up to occasional values higher than .60. It can be computed that if a 50-item test has a reliability of .90 and a validity of .40, the average item validity would be .16. Anyone who has seen an item analysis will know that item-test correlations vary widely in almost any set of items. It is clear then that item reliabilities also vary widely and it is probable that in most sets of items the item validities cover a wide range. It may be interesting to consider the comparative validity of a long test with the validity of a short test made up of somewhat more valid items. A test made

up of 200 items with an average item reliability of .15 and an average item validity of .12 would have a validity of .306. Extending such a test to infinite length would raise the validity to about .31. However 20 such items would have a validity of .27, and increasing the average validity of 20 items to value of .15 would result in a test validity of .34. Thus, a set of 20 good items may be more valid than a set of 200 items that are not so good. I have a strong suspicion that most sets of 200 items vary sufficiently in quality so that, for the majority of such lengthy tests, the best 20 items would be more valid than the total set.

A word of caution is clearly appropriate. When items are selected on the basis of item analysis data for use in a new test, sampling errors must be anticipated. If items have been selected because they are more homogeneous, more valid, more difficult, or easier than the items which have been rejected; then one must expect that the results of the new test will show a regression towards the mean of the original item pool. For example, if items are selected primarily for their validity data, the biggest effect of such regression will be the reduction of the item validities. There are those who believe that the effect of regression will generally be sufficient to invalidate the point I am hoping to establish. I doubt that. I doubt it particularly if statistics are used as they should be, as only a part of the item selection process. Item statistics should always be used to supplement editorial judgment, never to take its place. I do not know of sufficient experimental data on this problem, but I believe that item pools can reasonably be generated, and

many probably exist now, where item selection using item validity data will definitely enhance the validity of tests and permit the use of much shorter tests.

Data

There is sometimes a danger that conclusions based on theoretical or hypothetical data may be misleading because the mathematical model may neglect important practical considerations. Fortunately, I have had access to data from two quite different validity studies that may be used to evaluate these theoretical conclusions.

In one study the first 32 items of a 45-item multiple-choice test of writing ability had been validated against a carefully constructed criterion consisting of five essays. These essay tests were administered on separate occasions and each scored by five different readers. For a 300-case item-analysis sample the 32 items had a validity of .678. When the most valid 24 items were selected from this set, they had a validity of .716 for the item-analysis sample and .682 for the cross validation sample. By the way, the best eight items in this test had a validity of .566 for the cross validation sample. A good many 100- or 200-item tests do not achieve that level of validity.

The second study is quite different. Twelve liberal arts colleges or liberal arts divisions of universities participated in this study and the samples from these colleges ranged in size from 51 to 487 students for a total of 2,015 students. The criteria for this study were freshman year average grades normalized separately for each college on a 12-point scale. The test was a 150-item multiple-choice academic aptitude test that had

been prepared with more than usual care. The validity coefficients at these colleges for this test ranged from .39 to .64 and had an arithmetic mean of .5075. From this test a set of 48 items was selected and validated against freshmen average grades for about 80 percent of the total experimental group. The other students were dropped from this computation because they had not attempted all of the items. The validity coefficients for the 48-item test ranged from .42 to .67 and had a mean value of .5133, just a shade higher than the validity of the total test for the total group. In six colleges the shorter test had slightly lower validity than the total test; in four colleges the shorter test had a slightly higher validity, and in the two colleges with the largest samples the validity remained unchanged. These data tend to confirm my suspicion that lengthening a test beyond 50 items is not likely to increase its validity and they support the conclusion that it is better to invest in the best possible items rather than in the most possible items.

Discussion

Does all this really matter? Are the gains that may be achieved in validity big enough to justify questioning an honored tradition? Does this point of view represent a significant difference from what we have been doing all along? And finally, if we disregard reliability in order to enhance predictive validity, will we inadvertently weaken the content or the construct validities of our tests? These are important questions and they ought to be answered. They should be answered on the basis of research; but, until the research is done, they should be answered on the basis of careful shared opinion.

Obviously I believe that achievement tests in general are likely to be improved if more emphasis is placed on the importance of quality in items and less emphasis on the reliability of test score distributions.

The emphasis on quantity that has been a biproduct of our emphasis on reliability seems to have resulted in a Gresham's Law of items. We attempt to maximize the number of items we can include within fixed time limits and thereby allow very little time per item. As a result we tend to eliminate time consuming items, items that involve complex problems or items that require time for thought. Nedelsky (1965) classified achievement test items in three general categories: knowledge, understanding, and ability to learn. I think that the effect of this Gresham's Law has been to overemphasize the first category, knowledge, and to minimize the other two. Another possible classification of the content of a discipline has four categories: methods, observations, language, and theories. Again, I think that time pressure has resulted in overemphasis on the aspects of observations and language and underemphasis on theories and methods.

Reducing the time pressure in tests and allowing the candidates more time per item would, in my opinion, tend to improve the validity of achievement tests. Not only would it tend to permit more searching questions, questions calling for understanding and demonstrating the ability to learn, but also it might lead even farther. Knowing how to answer problems is not the highest form of knowledge. Knowing which problems are important problems and knowing how to define problems clearly are both more difficult and more valuable. This poses a serious challenge for standardized testing, but it is not a challenge that cannot be met and mastered. Once we understand that it is the quality of items,

not the quantity, that primarily determines the value and validity of an achievement test; we may be more willing to put our minds and budgets to this problem. Norman Frederiksen (1978) has made excellent progress in this aspect of test design, and there may be others who have done similarly effective work.

Conclusion

My first conclusion is this: adding to test reliability is not likely to add to test validity, and there is considerable danger that efforts to increase reliability may reduce validity. My second conclusion is that short tests may have useful validity if the items are well made and carefully selected. These conclusions lead to two policy recommendations. The first is to accept the obligation for occasional item validity studies. I say occasional because a worthwile item validity study is likely to be extremely expensive and cheap validity studies may well be misleading. My second recommendation is to put much more stress on item quality and less on quantity. A few good items are likely to be more valuable than an infinite number of less expensive items.

Tests are commonly built from existing pools of items, not from the hypothetical infinite pool sometimes referred to in test-score theory. The test makers' practical problem is to build the most valid test that can be built from a finite pool. In this common situation the appropriate strategy is to select only those items judged to have a high standard of validity and enough of them to adequately cover the content and constructs essential to the test. Adding more items, items with questionable validity, may increase the test reliability; but the additional items are likely to lower validity and make the test less useful than the shorter

test would have been. Reliability is not an adequate substitute for validity.

Postscript

The logic of this argument is incomplete. Test theory is written in the abstract language of mathematics, but the subject of test theory is not points in multidimensional space, but the responses of boys and girls, women and men, to test questions. Between the abstract notion of points in space and the practical problems of human test takers there stands a set of assumptions, approximations and simplifications. These assumptions are known to be less complicated than the reality they represent, although they may be sufficiently close to reality so that they can lead to useful insights into human nature and behavior, and useful insights into the characteristics of tests and how those characteristics influence the way they measure human nature.

One question that relates to both theoretical and practical concerns is whether maximizing validity for a particular criterion is likely to reduce the generality of that test's validity? Another question that may be important for this argument is to what extent and how commonly do departures from linear relationships have an important effect on the conclusions that can be derived from classical test theory? Latent trait theories such as the Rasch model anticipate curvilinear relationships between observed score and the abilities that tests are designed to measure. Such curvilinear relationships may be closer to actuality, particularly when scores pile up near the extremes of the possible score range, but latent trait theory has not yet had much to say about criteria and validities. The clearest and quickest answer to most of these questions will presumably come from experiments or quasi-experiments. The effort must be made to develop real criteria, criteria that would

clearly be preferred to the test score in practical decision situations. Reliability is not an adequate substitute for validity and test score theory is not an adequate substitute for experimental evidence. Until such evidence is available in greater quantity and quality, one must interpret theory with careful common sense.

This paper began with a hypothetical example made dramatic to give emphasis, but it dealt with a dichotomous situation. Another similarly dramatic example will describe a situation dealing with continuous variables. Consider the possibility that the members of a tennis team were staying at a hotel and that they became trapped together on an outside balcony by a fire. It appears to them that the only way of escape is to leap across to the roof of a nearby building and they have a rope, so they decide that their best jumper will go first with the rope to make a bridge for the others. Unfortunately they don't know which one of them is the best broadjumper and they don't have room to try on the balcony, but they do have room and time to try high jumping, assuming that the best jumper will be best at both kinds of jumping. Would the members of that tennis team be more interested in the reiability or the validity of their measurement?

References:

Frederiksen, Norman and Ward, William C., Measures for the Study of Creativity in Scientific Problem-Solving. Applied Psychological Measurement, 1978, 2, 1-24.

Gulliksen, Harold, Theory of Mental Tests. New York, Wiley, 1950

Nedelsky, Leo, Science Teaching and Testing. New York, Harcourt, Brace & World, 1965.